Charge to the work groups for Cochrane EBT project

Building on the first meeting at the 22nd Cochrane Colloquium in India, involving experts in toxicology, public health, animal testing, biostatistics, and systematic review methods, we propose the following work groups for developing a handbook for systematic review methods related to toxicology including nonhuman toxicology (NHT) and studies on mechanisms of toxicity. This document lays out areas where work is needed in order to develop this handbook.

Since our meeting, the National Toxicology Program of the US NIH has published a draft handbook for conducting a literature-based health assessment using its own approach for systematic review and evidence integration. This handbook covers many of the topics discussed below, but some major issues are not covered, such as the challenges in developing searches for accessing relevant studies. Dr Khris Thayer, the lead scientist on this effort, is a member of our workgroup, which will ensure appropriate integration of our efforts. We have also reached out to other leaders in this area.

Nonhuman toxicology and studies on mechanisms of toxicity are two of the three important domains of knowledge relevant to evaluating harms associated with environmental [and occupational] exposures; these are the sources of information that are currently relied upon by regulatory agencies and international agencies in reaching policy decisions concerning these exposures. Because of recommendations to incorporate systematic methods in this process, there is increasing interest in developing and validating systematic review methods for NHT. This handbook is limited to developing and validating methods for hazard identification, which is a qualitative finding of harm. We make this limitation because this is the first critical step in decision making in environmental and occupational health. There is considerable variation in policy approaches to the second step in decision making, which involves the quantitative assessment of risks [as a function of hazard and exposure]. This handbook accepts that the prevention of harms is the goal of decision making in environmental and occupational health as expressed by all national and international agencies; this goal introduces an emphasis on the application of study methods and statistical analyses that are validated as sensitive.

In this handbook, we do not propose to include methods for systematic reviews of human toxicology, which are almost entirely observational studies since ethical principles and guidelines do not accept randomized controlled studies in which human subjects are deliberately exposed to substances for the purpose of determining harms without expectation of benefit. There are several Cochrane Groups working on guidelines and methods related to evaluating evidence from observational studies, although mainly focused on the effectiveness of interventions rather than detecting harms. In the future we may expand our work group and seek a partnership with these existing Groups to develop an additional handbook on the evaluation of human studies. Eventual integration of human and non-human findings and conclusions will need to be considered, in light of their importance for decision makers and regulators.

Nonhuman toxicology (NHT) studies

There are two categories of experimental studies that utilize nonhuman models: toxicity testing (in which we include test guidelines and other regulatory procedures as well as research studies on specific toxic agents) and mechanistic studies. These are the only sources of information that can be used to provide evidence for actions to prevent harm prior to human exposure associated with environmental and occupational exposures as well as preclinical assessments of drug safety. Methods for systematic reviews of animal testing are available by Cochrane, SYRCLE and CAMARADES, but these methods primarily deal with the efficacy of the intervention rather than identification of harms. There are no existing validated methods for systematic reviews for the purpose of identifying harms.

GENERAL METHODS

We start with the general principles for Systematic Reviews that are fundamental to the Cochrane Collaboration, particularly a commitment to transparency in all steps of the process by which sources of bias are identified and evaluated, the acceptance of the need to validate all methods and to continuously improve them. Because of the lack of validated methods for NHT, testing different approaches will be important in this effort. At this stage, concerns about efficiency should not be paramount but will be an important element in the application of these methods. Existing case studies can be useful in this process: for example, the recent systematic reviews of arsenic (Maull et al. 2012; Navas-Acien et al. 2006) and PFOAS (Koustas et al. 2014). We aim to integrate, harmonize and embed in the Cochrane EBT handbook, whenever compatible with Cochrane guidelines, methods that were already adopted in Navigation Guide (http://prhe.ucsf.edu/prhe/navigationguide.html), NIEHS-OHAT Handbook for Systematic Reviews (http://ntp.niehs.nih.gov/pubhealth/hat/noms/index-2.html), SYRCLE (https://www.radboudumc.nl/Research/Organisationofresearch/Departments/cdl/SYRCLE/Pages/AboutSYRCLE.aspx), CAMARADES (http://www.dcn.ed.ac.uk/camarades/) and GRADE (http://www.gradeworkinggroup.org/intro.htm) and experts from all of these groups were engaged in the realization of the handbook.

Using the general template for systematic reviews as developed by the Cochrane Collaboration (http://handbook.cochrane.org/), we propose the following topics for the general methods, that correspond to the different workgroups:

1 Defining the review question and developing criteria for including studies

2 Searching for studies

3 Identifying eligible studies and collecting data

4 Assessing risk of bias in included studies

5 Analyzing data and undertaking meta-analyses

6 Addressing reporting biases

7 Presenting results

8 Interpreting results by preparation of 'summary of findings' tables and drawing conclusions

Defining the review question and developing criteria for including studies

a. Problem formulation for NHT:

The standard approaches of PECOT (population, exposure, comparison, outcome, time) can be applied to NHT with some redefinitions. Most importantly, since decision making in envir/occ health is related to control of exposures, NHT studies start with a defined exposure rather than outcomes or populations. In NHT, exposures can be defined as substances (and byproducts or metabolites) in the environment broadly defined to include environmental media as well as foods and consumer products. Substances may be specific [e.g., identifiable by a CAS number] or complex mixtures defined by source [e.g., coal fly ash; air pollution]. The definition of exposure usually includes the initiation and duration of exposure as well as the route(s) by which a human population is usually exposed; some of these may be similar in human and nonhuman models, but others are not [e.g., gavage for ingestion].

Populations in NHT are defined in detail to include the species, strain, sex, and age (or developmental stage) since these categories are relevant to design of studies as well as the evaluation of evidence in hazard identification. Other information may also be important, such as alterations in the genotype of the animals being studied, dietary restrictions, etc. Comparison groups in NHT are usually the same species being tested but not exposed to the substance of interest, similar to a placebo-controlled clinical trial. Amount and route of exposure are important aspects of the internal and external validity of a study and a major issue in assessing potential bias, especially when the goals of a systematic review are to evaluate evidence for harms at relatively low exposures or exposures that are intended to be relevant to those encountered by human populations. Exposures in NHT are often dissimilar to those expected or observed in human populations, a practice that is generally accepted as an important element in preventing harms [e.g., the use of higher exposures to elicit responses over the shorter life span of animals used in NHT]. *Timing*, both in terms of life stage and durations of exposure may emulate life stages in human populations, but the shorter lifespan of experimental animals is an important element in evaluation. Physiologically-based kinetic modeling (PBK) models are used to evaluate relevance in terms of internal dose, but these models are not fully validated for most substances.

The definition of *outcome* is one of the most challenging areas of problem formulation for NHT because of the challenges in evaluating the external validity of outcomes measurable in nonhuman organisms. In human studies, we can rely upon medical nosology to define specific outcomes within physiological systems, but no similar epistemic system exists for NHT. With the exception of cancer, most of the outcomes measurable in NHT have not been validated as

equivalent to human outcomes based on conserved mechanisms. For example, we have no validated animal analogues for diabetes; atherosclerosis; autism spectrum disorders and other neurocognitive and neuropsychiatric outcomes; most neurological diseases; and autoimmune diseases. Given these limitations, there are at least three options for problem formulation and protocol development in terms of defining NHT outcomes relevant to human disease:

- Rely upon key words and textual analyses [of abstracts] to determine that the study is relevant to a disease outcome observed in humans (unknown external validity –trash in, trash out)
- Compile a list of externally valid non-human models and outcomes (including biomarkers) by the reviewing group proven to be relevant to a defined disease outcome observed in humans, such as measurements of biomarkers, histopathology, dysfunction, or other observations.
- Develop a hybrid approach that includes citation of human disease outcomes and relevant terms for limiting reporting bias, but including only externally valid nonhuman models in the following study selection steps

b. Problem formulation for mechanistic studies:

This body of information is derived from studies most of which have uncertain external validity [i.e., demonstrable relevance to human health]. Genotoxicity (assessed through in vivo micronuclei/COMET test by National Toxicology Program), is a well-defined and relevant property of carcinogens, and a good example of externally valid and validated early marker for potential carcinogenicity, because of its well defined mechanism of action that is known to be conserved phylogenetically from eukaryotic cells to humans. These conditions have not been met for mechanistic studies related to noncancer harms. In many cases, we are not certain of the relevant mechanisms in humans, and in other cases, research indicates that multiple mechanisms may be involved in the steps from exposure to outcome. In terms of internal validity for mechanistic studies, there are some efforts underway to develop criteria by which mechanistic studies can be evaluated [such as at NIEHS and US EPA]. In the absence of this information, guidelines for problem formulation can only be tentative and testing proposed guidelines is a priority for this handbook. The development of *a priori* criteria for exclusion and inclusion of mechanistic studies is an important part of the work of our group at this stage.

Searching for studies

After formulating the problem and defining sources of primary studies, the process of searching for NHT studies will require considerable resources of expert informaticists and toxicologists. There have been examples of literature searches in several systematic reviews that have included NHT (Krauth et al. 2013; Maull et al. 2012; Navas-Acien et al. 2006).There are no published reports on validation of these protocols. There remain challenges in searching the literature for NHT related to the issues discussed above, especially, but not limited to, defining outcomes. As a result, the replicability of searches is often poor and the failure to

access the same set of relevant studies is a major contributor to variations in health assessments among researchers and regulatory agencies.

Identifying eligible studies and collecting data

Some general principles, such as sources of studies and avoidance of duplication, can be applied from CC experience. The NAVIGATION project has developed material on this topic for NHT [reference (Woodruff and Sutton 2014)]. These require validation. For mechanistic studies, there is a lack of criteria development aside from adoption of good laboratory practices and other guidelines for annotating methods used in studies involving cell culture, gene expression analyses, and metagenomics. Work is needed in this area. Some practitioners advocate broad inclusion, far beyond what is generally acceptable in EBM/HC, but reflective of the sources of some of the information in NHT which are not limited to the peer reviewed literature.

Methods on data extraction are under development (Rooney et al. 2014). However, they are as yet incomplete as we have not developed a consensus on the data to be extracted from each study. Data should be sufficient to support analyses of confidence in each study and risk of bias.

Assessing risk of bias in included studies

Many of the elements related to assessing risks of bias that have been identified for CC systematic reviews of interventions are relevant here, with the additional concerns for appropriateness of study design. We stress that GLP guidelines do not provide sufficient guidance on the internal and external validity of these studies. The elements of GLP guidelines need to be validated for their relevance to NHT. In addition to the concerns for the external validity of the study design (with respect to the human health event of concern, and with respect to the endpoint being assessed in the study), there are issues related to the internal validity of methods used in NHT and particularly the identification of appropriate statistical analyses for each study under evaluation the study (including the *a priori* power calculation for the study design). Inclusion of a known "positive" substance [that is an agent recognized to cause a specific outcome] is often recommended as a test of internal validity (that is, does the study measure what it is purported to measure), but this is not always possible.

With respect to external validity (that is, does the study produce information that is relevant to human health), the lack of information on this topic constitutes a major potential source of bias and uncertainty. Assessing bias and confidence in mechanistic studies is challenging because of even greater uncertainties as to external validity. Moreover, there is a broad range to study designs – including *in silico* computer modeling – used in these studies, which affects our current ability to integrate the results of these studies as a group. Proposals, such as matrix read-across databases, have not been validated (Silbergeld et al. 2015).

COI is an important criterion for confidence since there is evidence that funding sources are associated with predictable outcomes of NHT research. In light of the recent evidence supporting the inclusion of funding source as a standard item for risk of bias in Cochrane and considering how ubiquitous are the conflicts of interest in the field of toxicology, inducing several Institutions to implement more stringent criteria for addressing Conflict of Interest (for example NTP, EPA and EFSA), we suggest to include and validate funding source as an item in the risk of bias assessment of NHT from the inception of this handbook.

Data integration and meta-analyses

Meta-analyses of NHT and mechanistic studies present several different challenges compared to meta-analyses of interventions, in particular since the goals of NHT studies are to detect harms and not benefits (absence or presence of positive effects is not the aim of a toxicological evaluation). In addition, there is a need to develop guidelines for assessing the appropriateness of a meta-analysis in NHT since the literature will often be highly heterogeneous in terms of different populations (phyla, species and strains), exposures, timing, outcome definitions, and methods of measurement. It may be a default to suggest that different species cannot be integrated in the same meta-analysis.

Addressing reporting biases

Reporting bias arises when the dissemination of research findings is influenced by the nature and direction of results. The phenomenon is scarcely addressed in the toxicological literature, although several preliminary evidences suggest that this may be often present (Weuve et al. 2012), particularly in light of the great amount of information not accessible for the public and often exclusively available to governments and/or industry (Gee et al. 2013). Similarly to the medical field, at least three domains have to be explored : publication bias (unpublished studies tend to have different results than published studies), selective outcome reporting (statistically significant outcomes have a higher likelihood of being reported compared to non-significant findings) ascertainment bias (studies that are easier to find may have different results from those that are harder to find and that belong to the so-called grey literature (Bolland and Grey 2014). Funnel plots can be used for reviews with sufficient numbers of included studies, but an asymmetrical funnel plot should not be equated with publication bias.

Presenting results and 'Summary of findings' tables

Study flow diagrams are used to illustrate the results of the search and the process of screening and selecting studies for inclusion in the review. A flow diagram using the PRISMA-like template for NHT is one method that can be supported by the text of the flow diagram, distinguishing studies and records. Records are information sources about a study, such as journal articles,

book chapters, web pages and other documents. Studies are the research enterprises themselves, in this case the primary studies in non-human human toxicology. Usually a flow diagram will start by describing numbers of records retrieved (the majority of which will typically be from bibliographic databases). Following de-duplication, the records will have been mapped onto distinct studies and the flow diagram will reflect this by switching its emphasis to studies.

The flow diagram should present:

- number of unique records identified by the searches;
- number of records excluded after preliminary screening (e.g. of titles and abstracts);
- number of records retrieved in full text;
- number of records or studies excluded after assessment of the full text, with brief reasons;
- number of studies meeting eligibility criteria for the review (and thus contributing to qualitative synthesis); and
- number of studies contributing to the main outcome.

Integrating results and drawing conclusions

Integration of study results should be guided by the goal of identifying harms, that is, to ensure safety. For that reason, the standard methods used in integrating data from studies of interventions (clinical trials or observational epidemiology] may need reconsideration in NHT. The use of meta-analytic methods, either quantitative or narrative, should not be conducted in a manner that weighs the direction and magnitude of effect of all included studies that meet criteria for evaluation, since meta-analytic methods that integrate magnitude and direction of effect across studies may be misleading. The finding of hazard in one set of studies using a specific study design or population is often of importance in decision making because of differences in study design as fundamental as the species tested (e.g., thalidomide experience). Experience in NHT indicates that strain as well as species can be reliably predictive of responsiveness to an exposure, such that combining different strains of rats, for example, may introduce systematic bias in terms of studies on species that are insensitive or nonresponsive to certain substances or in terms of specific outcomes [such as cancer, or arsenic]. These evaluations require expertise in NHT.

Some elements that have been proposed are of debatable value in terms of their relationship to confidence in study results: the size of an effect or severity/prevalence of the outcome is not relevant to hazard identification since this is highly dependent upon the exposure utilized; consistency across studies must be tempered by considerations of the heterogeneity among studies as discussed above.

Thus, one of the main issues for this topic is the extent to which NHT studies can be combined without accumulating excessive heterogeneity. This concern is raised because of proposals to

integrate NHT studies testing different specific outcomes, if each outcome is plausibly relevant to the human disease of concern (e.g., studies on insulin resistance, glycosylated hemoglobin, glucose levels, and organ pathology as all indicative of diabetes; or studies on maze performance, conditioned response tests, and short term memory as all indicative of neurocognitive deficits). A narrative summary may be undertaken, but more formal integration requires validation of methods.

Integration of mechanistic evidence with NHT studies is also not recommended for the reasons discussed above related to the validation of mechanistic studies (external and internal validity). The availability of mechanistic evidence may increase confidence, but the lack of such evidence cannot decrease confidence.

Work is currently underway on the applicability of GRADE as a rubric for the qualitative consideration of studies. One issue to consider is the starting point: Guyatt has argued that for observational epidemiology studies, the initial grade is set low, to indicate the inherently lower reliability of these designs as compared to controlled clinical trials. This initial grade can be increased depending upon evidence of adequate control of bias in each study. This may be a useful strategy for NHT as well.

'Summary of findings' tables present the main findings of a review in a transparent and simple tabular format. In particular, they provide key information concerning the quality of evidence, the hazard and the sum of available data on the main outcomes. Most reviews would be expected to have a single 'Summary of findings' table. Other reviews may include more than one, for example if the review addresses more than one hazard. Implementation of a specific and validated GRADE system for NHT is a fundamental on-going aspect for this chapter.

Note: the nature of integration depends in part upon the use of NHT evidence in decision making or the policy structure within which the systematic review is undertaken. The IARC rubric and the NTP rubric (which are similar) will be supported by a different integration strategy than a rubric of risk assessment or the Precautionary Principle.

REFERENCES

Bolland MJ, Grey A. 2014. A comparison of adverse event and fracture efficacy data for strontium ranelate in regulatory documents and the publication record. BMJ open 4:e005787.
Gee D, Grandjean P, Hansen SF, van denHove S, MacGarvin M, Martin J, et al. 2013. Late lessons from early warnings: Science, precaution, innovation.European Environment Agency.
Koustas E, Lam J, Sutton P, Johnson PI, Atchley DS, Sen S, et al. 2014. The navigation guide - evidence-based medicine meets environmental health: Systematic review of nonhuman evidence for pfoa effects on fetal growth. Environmental health perspectives 122:1015-1027.
Krauth D, Woodruff TJ, Bero L. 2013. Instruments for assessing risk of bias and other methodological criteria of published animal studies: A systematic review. Environmental health perspectives 121:985-992.
Maull EA, Ahsan H, Edwards J, Longnecker MP, Navas-Acien A, Pi J, et al. 2012. Evaluation of the association between arsenic and diabetes: A national toxicology program workshop review. Environmental health perspectives 120:1658-1670.
Navas-Acien A, Silbergeld EK, Streeter RA, Clark JM, Burke TA, Guallar E. 2006. Arsenic exposure and type 2 diabetes: A systematic review of the experimental and epidemiological evidence. Environmental health perspectives 114:641-648.
Rooney AA, Boyles AL, Wolfe MS, Bucher JR, Thayer KA. 2014. Systematic review and evidence integration for literature-based environmental health science assessments. Environmental health perspectives 122:711-718.
Silbergeld EK, Mandrioli D, Cranor CF. 2015. Regulating chemicals: Law, science, and the unbearable burdens of regulation. Annual Review of Public Health 36:null.
Weuve J, Tchetgen Tchetgen EJ, Glymour MM, Beck TL, Aggarwal NT, Wilson RS, et al. 2012. Accounting for bias due to selective attrition: The example of smoking and cognitive decline. Epidemiology (Cambridge, Mass) 23:119-128.
Woodruff TJ, Sutton P. 2014. The navigation guide systematic review methodology: A rigorous and transparent method for translating environmental health science into better health outcomes. Environmental health perspectives 122:1007-1014.